This document contains 6 practical assignments, oriented on real data. The data are available in Moodle. Solutions to the first two assignments are available in sections 4.5.2 and 4.5.3 of Leo Belzile's tutorials.

Working through any of the practical tasks will require these steps:

1. loading the data – download a file from moodle and check commands such as `load()` or `read.table()`; for the `pima` dataset, one can just load the `faraway` package,
2. understanding the data – probably requires some googling,
3. exploring the data – making a first impression about the dependencies, looking for potential outliers or other problems
4. building a model or several candidate models – manually using model-submodel tests or in an automated manner,
5. critical reflection on your model(s) – are the assumptions satisfied? can the model be improved?

Often, data are collected with a particular scientific question in mind. This question also drives the model selection, which is why manual model building is often preferable to automated model selection methods such as cross-validation. Also note that different number of observations (n) allow for models of different complexities. Finally, potential outliers cannot be left out at the beginning based merely on an analyst's hunch, their influence on the final model (used to answer the scientific questions) has to be analyzed thoroughly when no expert knowledge allowing to discard them is available.

Practical 1

The dataset `windmill` contains measurements of electricity output of wind turbine over 25 separate fifteen minute periods. We are interested in the relation between direct output and the average wind speed (measured in miles per hour) during the recording. To load the dataset, use the commands

1. Fit a linear model with wind speed as covariate and plot the standardized residuals against the fitted values. Do you notice any residual structure missed by the model mean specification? Try fitting a model using the reciprocal of wind speed as covariate. Comment on the adequacy of the models.
2. Predict, using both models in term, the output of electricity given that the average wind speed in a given period is 5 miles per hour. Provide prediction intervals for your estimates.
3. The electricity production should be zero if there is no wind, yet this is not captured by the model linking output to velocity. Update your model to remove the intercept. What are the consequences of the latter? Comment on the impact on your goodness-of-fit diagnostics and the standard errors of the estimated effect for velocity.
4. Produce a standard Gaussian quantile-quantile plot of the residuals obtained from the orthogonal transformation described in the midterm (normalize the residuals by first subtracting their mean and then dividing them by their standard deviation).

Compiled 09/09/2019 at 14:07

Practical 2

The time series `AirPassengers` provides the monthly totals of international airline passengers from 1949 to 1960.

1. Fit a linear model linking the number of passengers to the time of the year. Consider adding a monthly effect using dummies. Do you notice a significant improvement in fit? Use the model to predict the number of passengers in December 1962.
2. Provide diagnostic plots (the method `plot` returns six graphs for objects of class `lm`). What do you notice?
3. There may be evidence that the growth in aerial traffic is exponential during the time period under study. Try fitting a linear model with the log of the number of passengers as response. Produce diagnostic plots and make a quantile-quantile plot of the externally studentized residuals (the function `rstudent` applied on an object of class `lm` returns the latter).
4. Plot the lagged residuals, i.e. plot $\boldsymbol{e}_{-1}$ against $\boldsymbol{e}_{-n}$. Is any of the hypothesis of the linear model seemingly violated?

Practical 3

Consider data set pima, which can be found in faraway package in R. We are interested in modeling the body mass index bmi of female Pima Indians depending on the number of pregnancies pregnant.

1. Remove observations for which bmi is zero from the data set (these are likely errors).
2. Treat pregnant as factor variable, but consider binning the variable into o6 categories: 0,1,2,3-4,5-6,7 or more pregnancies.
3. Explore other measured variables and decide whether they have significant impact on the relationship between bmi and pregnant. Build a linear regression model.
4. Using your model, decide by a statistical test whether bmi depends on pregnant or not.

Compiled 09/09/2019 at 14:07

Practical 4

The wages data set is extracted from the book

> E. R. Berndt. *The practice of econometrics: classic and contemporary*, 1991, Addison-Wesley Pub. Co., 702 p.

The description accompanying the data set is reproduced here for convenience:

> *This data set consists of 534 randomly selected employed workers from the May 1985 current population survey conducted by the U.S. Department of Commerce. This is a survey of over 50,000 households conducted monthly, and it serves as the basis for the national employment and unemployment statistics. Data are collected on a number of individual characteristics as well as employment status.*

The data set contains the following variables:

`ED`	years of education
`SOUTH`	1 if lives in south
`NONWH`	1 if nonwhite
`HISP`	1 if Hispanic
`FE`	1 if female
`MARR`	1 if married with spouse present (in household)
`MARRFE`	1 if married female with spouse present
`EX`	years of labor market experience (`AGE-ED-6`) (minimum = 0 imposed ex post)
`EXSQ`	years of labor market experience squared
`UNION`	1 if working on a union job
`LNWAGE`	natural logarithm of average hourly earnings
`AGE`	age in years
`MANUF`	1 if working in manufacturing industry
`CONSTR`	1 if working in construction industry
`MANAG`	1 if occupation is managerial or administrative
`SALES`	1 if occupation is sales worker
`CLER`	1 if occupation is clerical worker
`SERV`	1 if occupation is service worker
`PROF`	1 if occupation is professional/technical worker

Build a linear model, which can answer the following questions:

1. Quantify, if any, the net effects of education, work experience and union membership on hourly wages.
2. Assess the existence of a gender gap, accounting for the presence of confounding variables.

Compiled 09/09/2019 at 14:07

Practical 5

This data set was extracted from

> Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989) *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients*, Journal of Urology 141(5), 1076–1083.

This data set is described in Wakefield (2013), pp. 5-6.

> *The data were collected on $n = 97$ men before radical prostatectomy, a major surgical operation that removes the entire prostate gland along with some surrounding tissue. [...] In Stamey et al. (1989), prostate specific antigen (PSA) was proposed as a preoperative marker to predict the clinical stage of cancer. As well as modeling the stage of cancer as a function of PSA, the authors also examined PSA as a function of age and seven other histological and morphometric covariates. [...] The BPH and capsular penetration variables originally contained zeros, and a small number was substituted before the log transform was taken. It is not clear from the original paper why the log transform was taken though PSA varies over a wide range, and so linearity of the mean model may be aided by the log transform. It is also not clear why the variable PGS45 was constructed.*

The data set contains the following variables:

`lcavol`	log of cancer volume, measured in milliliters (cc). The area of cancer was measured from digitized images and multiplied by a thickness to produce a volume.
`lweight`	log of the prostate weight, measured in grams.
`age`	The age of the patient, in years.
`lbph`	log of the amount of benign prostatic hyperplasia (BPH), a noncancerous enlargement of the prostate gland, as an area in a digitized image and reported in cm^2.
`svi`	seminal vesicle invasion, a 0/1 indicator of whether prostate cancer cells have invaded the seminal vesicle.
`lcp`	log of the capsular penetration, which represents the level of extension of cancer into the capsule (the fibrous tissue which acts as an outer lining of the prostate gland), measured as the linear extent of penetration, in cm.
`gleason`	Gleason score, a measure of the degree of aggressiveness of the tumor. The Gleason grading system assigns a grade (1–5) to each of the two largest areas of cancer in the tissue samples with 1 being the least aggressive and 5 the most aggressive; the two grades are then added together to produce the Gleason score.
`pgg45`	percentage of Gleason scores that are 4 or 5.
`lpsa`	log of prostate specific antigen (PSA), a concentration measured in ng/m

Answer the following questions: questions:

1. Quantify changes, if any, in PSA that are due to interactions between cancer volume and the other explanatory variables.

2. Assess whether the PSA is a useful marker of prostate cancer volume and the limitations of your model in assessing this.

Practical 6

The Oxford Parkinson's Disease Telemonitoring data set is extracted from the website

> Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). Irvine, CA: University of California, School of Information and Computer Science.

and was used in

> Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig (2010), *Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests*. IEEE Transactions on Biomedical Engineering, **57**(4).

The goal of the study was to use voice symptoms severity recorded remotely in patients houses and use the later to predict a diagnostic measure of Parkinson's disease. The use of vocal impairment measures is due to its prevalence among patients with Parkinson's disease at early stages and to evidences that it is an indicator of the progress of the disease.

The data set contains multiple recordings from each of 42 patients in the study, for a total of $n = 5875$ recordings. Relevant excerpts of the paper are reproduced below for convenience:

> *Physical test observations are mapped to a metric specifically designed to follow disease progression, typically the unified Parkinson's disease rating scale (UPDRS) that reflects the pres- ence and severity of symptoms (but does not quantify their underlying causes). For untreated patients, it spans the range 0–176, with 0 representing healthy state and 176 representing total disabilities, and consists of three sections: 1) mentation, behavior, and mood; 2) activities of daily living; and 3) motor. The motor UPDRS ranges from 0 to 108, with 0 denoting symptom free and 108 denoting severe motor impairment, and en- compasses tasks such as speech, facial expression, tremor, and rigidity. Speech has two explicit headings and ranges between 0 and 8 with 8 being unintelligible. [...] UPDRS values were obtained at baseline, three-month and six-month trial periods, but the voice recordings were obtained at weekly intervals*

Most explanatory variables are dysphonia measures obtained from the KayPentax multidimensional voice program (KP-MDVP) and preprocessed using a software; refer to the paper for a description of their signification (section C).

`UPDRSmotor`	clinician's motor UPDRS score, linearly interpolated
`time`	time since recruitment into the trial (in days)
`id`	subject identifier
`age`	subject age (in years)
`sex`	1 if female
`jitter`	KP-MDVP absolute jitter (in microseconds)
`RAP`	KP-MDVP relative amplitude perturbation
`PPQ5`	KP-MDVP five point period perturbation quotient
`DDP`	KP-MDVP average absolute difference between cycles, divided by the average period
`shimmer`	KP-MDVP local shimmer (in decibels)
`APQ3`	KP-MDVP three point amplitude perturbation quotient
`APQ5`	KP-MDVP five point amplitude perturbation quotient
`APQ11`	KP-MDVP 11 point amplitude perturbation quotient
`DDA`	average absolute difference between consecutive differences between the amplitudes of consecutive periods
`NHR`	noise-to-harmonics ratio for tonal components in the voice
`HNR`	harmonics-to-noise ratio for tonal components in the voice
`RPDE`	recurrence period density entropy, nonlinear dynamical complexity measure
`DFA`	detrended fluctuation analysis (signal fractal scaling exponent)
`PPE`	pitch period entropy, a nonlinear measure of fundamental frequency variation

The dysphonia measures can be broadly categorized as follows:

- `jitter`, `RAP`, `PPQ5` and `DDP` are measures of cycle-to-cycle variability of the fundamental frequency, F_0.
- `shimmer`, `APQ3`, `APQ5`, `APQ11`, `DDA` are measures of the amplitude variability in F_0.
- `NHR`, `HNR`, `RPDE`, `DFA`, `PPE` are measure of variation in voice amplitude.

Discuss whether vocal recordings could be used to adequately predict motor UPDRS. Obtain a parsimonious model for the motor UPDRS and use the latter to determine whether there were any differences in the predictive performance, on average, between patients at different stages of the disease based on their total UPDRS score.